How do Genomes Differ?

Genomes

When considering genomes and their products, many genes involved in basic processes that are vital to life are highly conserved across all organisms. For example, enzymes involved in metabolic pathways are similar from bacteria to humans – functionally they have to interact with the same substrates and products, catalyse the same reactions, and are subject to constraints during evolution. This makes it possible to carry out experiments in the laboratory setting: for any human gene, or pathway, of interest to the cell biologist or molecular geneticist there will be a suitable model organism, or cell system, to test out a hypothesis.

Between different mammals, genomes are largely conserved in gene content, although species specific expansions of DNA sequences and some novel or highly diverging sequence signatures may also be observed. Techniques such as chromosome painting have confirmed that large blocks of conservation (syntenies) exist between both closely related and distant mammalian species. Most of the differences we observe both between and within species are dependent upon subtle variation in the spatial, temporal and quantitative expression of the genomic products.

Aims:

- Consider normal polymorphism in the genome and the impact that this has on phenotypic variation between individuals in a population.
- Look at different types of polymorphism and how they are assessed.
- Why polymorphic markers can be used to find locations of disease genes.
- Introduce ideas and concepts needed or future lectures.

You need to have background knowledge of gene transcription and translation. Please refer to any good molecular biology textbook as a refresher.

Understanding the importance of natural variation in the genome can help us to understand the pathogenic changes to DNA sequence or chromatin modifications through epigenetic mechanisms.

Types of variation:

At the level of DNA in the genome.

There are two major types of tandem repeat

1. Minisatellites (VNTR): 6-100bp units that are repetitive. Tend to be polarised towards the telomeres and centromeres of chromosomes. Highly polymorphic and good for forensic studies/ paternity testing etc. Also called VNTRs (variable number tandem repeats).
2. Microsatellites (MS, STR, SSR): small (1-5bp) repetitive units scattered throughout the genome at a reasonably regular spacing (about 1 per 100Kb). Easy to determine through PCR based analyses. Highly polymorphic. Good for genome analysis. Used for physical mapping, genetic mapping, family studies to establish linkage, association studies. Also called short tandem repeats (STR) and short sequence repeats (SSR).

The commonest changes are the ***S***ingle ***N***ucleotide ***P***olymorphisms (SNPs), where the sequence is altered: e.g. AT***C***GT → AT***T***GT.

The change can either be a transition (C←→T, G←→A) or transversion (a purine to a pyrimidine).

These changes mainly arise as errors during DNA replication (estimates predict each new diploid genome probably has 175 novel SNPs that arise mostly as the result of replication errors: real experimental evidence suggests the rate is about half of this), or as a consequence of exposure to environmental and chemical carcinogens (mutagenesis).

Most DNA variants are neutral and tolerated from one generation to the next when they are present in the germ line.

Somatic mutations also occur and can contribute to the changes observed in cancer cells with respect to normal tissue from the same individual.

CpG dinucleotides that are subject to methylation in the genome (^{m}CpGs) are also particularly sensitive to transition events, as the methylated C can be spontaneously deaminated as a consequence of a tautomeric shift to generate a T (there is also enzymatic deamination in some species of bacteria). Some genetic conditions occur at high *de novo* frequency in the population owing to the occurrence of 'hotspots' for mutation in ^{m}CpGs. This was first noted in some of the pedigrees with haemophilia, where the same bases within the X-linked factor IX gene showed mutation in unrelated individuals.
(*Think about recurrence of somatic mutation at specific sites in cancer biology as well. Here there is a* ***selection*** *of the favourable mutations, since they provide a growth advantage to the cell. In analysis of tumours v normal tissues, the same base pair changes are observed in different individuals.*)

Some minor sequence differences are due to small insertions and deletions (indels). Often, these are only a few base pairs in length, and if they occur in a coding sequence, a multiple of 3 bp (i.e. a codon) will maintain the open reading frame. A common Cystic Fibrosis mutation is a deletion of a single amino acid – however, genetic disease is not a common theme of such codon length indels. Normally, frame-shifts are caused owing to a single or two base pair frame shift – see introduction of novel polypeptide sequence and termination codons.

CNVs (copy number variants) are much larger regions of DNA that are subject to duplication/ deletion. Many of the mapped CNVs are found in normal individuals. They are evolutionarily important as there is evidence of gene duplication followed by sequence divergence as a way to generate gene families (e.g. Hb loci, MHC loci). Most CNVs are flanked by sequences showing a high level of sequence homology, suggesting that a non-allelic form of recombination is the mechanism by which these genome duplications arise. Again, some regions of the genome are more susceptible and explain the higher rate of *de novo* occurrence of some genetic disorders where gene dosage is critical to the normal phenotype. The

DECIPHER project looks at integrating the data from multiple pedigrees to help distinguish which CNVs may be pathogenic rather than polymorphic. Pathways most readily disrupted through CNVs are those with tight regulation of gene expression. The control of foetal growth and brain development appears to be amongst these, as copy number perturbation (e.g. through unbalanced translocations, imprinting defects, and intra-chromosomal large CNVs) is often associated with growth defects *in utero*, and with both non-specific and syndromal MR/ behavioural outcomes.

Reading list for CNV studies - a few suggestions

1: Swaminathan GJ, Bragin E, Chatzimichali EA, Corpas M, Bevan AP, Wright CF, Carter NP, Hurles ME, Firth HV. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. Hum Mol Genet. 2012 Oct 15;21(R1):R37-44. Epub 2012 Sep 8. PubMed PMID: 22962312; PubMed Central PMCID: PMC3459644.

2: Rucker JJ, McGuffin P. Genomic structural variation in psychiatric disorders. Dev Psychopathol. 2012 Nov;24(4):1335-44. doi: 10.1017/S0954579412000740. PubMed PMID: 23062301.

3: Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. Curr Opin Genet Dev. 2012 Jun;22(3):229-37. doi: 10.1016/j.gde.2012.03.002. Epub 2012 Mar 29. Review. PubMed PMID: 22463983.

4: Southard AE, Edelmann LJ, Gelb BD. Role of copy number variants in structural birth defects. Pediatrics. 2012 Apr;129(4):755-63. Epub 2012 Mar 19. Review. PubMed PMID: 22430448.

5: Klopocki E, Mundlos S. Copy-number variations, noncoding sequences, and human phenotypes. Annu Rev Genomics Hum Genet. 2011 Sep 22;12:53-72. Review. PubMed PMID: 21756107.

Changes not affecting DNA sequence

Methylation of the genome is normal: for example, in males there is only one X chromosome, so in females, one of the X chromosomes is methylated to help ensure gene dosage is similar between XX and XY individuals. Such X-inactivation is the most extreme case.
Other regions of the genome may also undergo DNA modification to help control levels of transcription. Often these modifications at the base level occur in conjunction with histone modification (see below). Imprinting at autosomal loci ensures mono-allelic expression, with either the maternal or paternal allele active (this is locus dependent, and the allele expressed will be consistent across a species e.g. always maternal).

Chromosomes consist of both DNA and protein: while the above sections refer to variation at the DNA level, histones are also subject to modification (e.g. methylation and acetylation). Again, modification of the histones helps to regulate expression of nearby genes. Both nucleic acid and protein modification of the chromosome are important in the study of parental imprinting and of epigenetics (phenotypic outcome **not** dependent on DNA sequence variation). This is an increasingly active area of research both in investigating the somatic changes involved in cancer genetics, and as we become aware that DNA variation is often unable to fully resolve the riddle of complex disease!
There is some evidence that differences in the epigenetic modifications might provide a link between environment-gene interaction, or even penetrance of disease loci: e.g. twin studies show that the somatic changes to DNA/ histone modification that occur over time may not occur at the same rate in monozygotic twins.

At the level of the chromosome: contiguous gene deletion syndromes, imprinting, unbalanced translocations
Although some of the common CNVs described in the literature are quite large, and may harbour many genes, alteration in gene copy number in these cases appears to be neutral (note CNVs may result in 3 copies v 2, or 1 copy v 2, depending on gain/loss of material). It is also true that many genes are perfectly sufficient at haploid dosage. Many single-gene KO mice

have been generated, but only a small number result in an overt phenotype that can be ascribed to haploinsufficiency. In reality, they do not cause an outcome outside of the ‘normal’ range.

This is not true for every gene - chromosomal deletion syndromes or unbalanced translocations that alter the allelic balance between large numbers of genes, are more likely to disrupt key pathways, thus leading to a phenotype. Note too that some genes in large deletions may carry a parental imprint, where the origin of the active copy of the locus is important (later lectures in this option). We understand some of these in detail, but the majority of imprinted genes are relatively poorly studied to date.

Most of these large chromosomal abnormalities arise through errors of chromosome pairing and segregation in germ cell maturation (i.e. the chromosomal aneuploidies and translocation events).

Variation: Summary

Most changes at the DNA level appear to be tolerated: if they were not we would have no genetic markers for our studies! The majority will be ‘neutral’. This is partly because

1. Most of the DNA does not encode protein.
2. The genetic code is degenerate, so not all changes in a protein coding sequence will alter the coding potential of the open reading frame.
3. Some amino acids can be interchanged (conservative replacements) without substantially altering protein function.

Some changes to gene sequences may result in a non-viable embryo (a non-viable zygote, or, a zygote able to proceed to a specific developmental stage followed by embryonic lethality), or cause infertility (non-viable germ cell, or failure of somatic cells to support germ cell development). In these cases, we do not observe an outcome in our population, the change is not

passed on to the next generation, and the variant is rapidly lost from the polymorphism pool.

Other apparently deleterious changes (the real ‘genetic mutations’ that result in disease) persist in populations in spite of the impact on the viability of the individual. These are often mutations that are life-threatening if homozygous, but offer some advantage to a heterozygous carrier. Classic examples are the mutations responsible for sickle cell anaemia, and cystic fibrosis. These are the most overt examples of SNPs and small indels that are present in different human populations as a consequence of positive selection by the endemic diseases.

DNA variation with an impact on phenotype will be in regions of the genome that include open reading frames, or in regulatory elements, such as promoters, enhancers and ncRNA. All these can alter gene expression. Often literature reviews will refer to these as ‘functional polymorphisms’.

Most DNA variants result in phenotypes that are considered within the normal range. **BUT**
Genetic variation interacts with the environment, so changes to lifestyle, drug development for disease treatment, novel or altered pathogen exposures may all reveal new phenotypic outcomes that would not be evident in our ancestors who also carried these variants.

Using Variation

1. Genetic maps: places loci in relative order and distance based on recombination. Sex-limited or sex-averaged maps can be developed for all autosomes and the X. ***The Y cannot be mapped genetically (nor can the mitochondrial genome).***
2. Linkage analysis in pedigrees to identify genomic regions and loci responsible for disease phenotypes. Both MS and SNPs are used.
3. Association analysis at the population level mainly uses SNPs as these are easily typed using microarray based technology (although older

studies have used MS for a first phase pass of the human genome). Association always goes with a specific allele at the locus of interest!

4. Epigenetic studies use a combination of techniques to determine differences in chromatin modification. Chemical modifications of DNA allow the detection of sequences with ^{m}C through either direct sequencing or array based screening. Antibodies developed against modified histones allow the capture of regions of modified chromatin from the genome. This is followed by isolation and identification of the associated DNA sequence to pinpoint the loci that are marked by histone modification. Again this last part can be array based.

Much of the above will be covered in more depth in your other lectures this year.

Suggestions for further reading

Refresh your memory of gene translation and transcription with any good molecular biology textbook!

All papers are available on-line in the @cam domain

McClellan J and King M-C. 2010, Cell 141: 210-217 Genetic Heterogeneity in Human Disease. *A good review to get you thinking about this lecture series!*

Tabor HK, Risch NJ, Myers RM. 2002, Nature Reviews (Genetics) 3:391-7. Candidate gene approaches for studying complex traits: practical considerations. *A good paper to read within context of genetic mapping approaches and where it goes next to identify key genes. Nice summary of types of mutations in genes.*

Arnheim N, Calabrese P. 2009 Nat Rev Genet.10:478-88. Understanding what determines the frequency and pattern of human germline mutations. *Good overview of how and why germline mutations arise, and also the impact of gender and age.*

Skelly DA, Ronald J, Akey JM. 2009 Annu Rev Genomics Hum Genet.10:313-32. Inherited variation in gene expression. *A thorough summary of how variation has an impact on gene expression.*

Nordstrand et al., 2007, Neuroscience, 145:1309-1317. Genome instability and DNA Damage accumulation in gene-targeted mice. *Considers different mechanisms that can introduce mutations to DNA sequence*.

Illingworth and Bird: 2009, FEBs letters, 583: 1713-1720. CpG Islands – 'A rough guide'. *Good introduction to the importance of CpG based promoters*.

Nakamura Y, Journal of Human Genetics (2009) 54, 1–8 DNA variations in human and medical genetics: 25 years of my experience. *A historical perspective on genetics, genomics and the relevance of genetic variation.*

Applications of SNPs to dissect the role of genome variation in disease

Modern Approaches

Regardless of whether we are looking at single genes to define mutations in monogenic disorders, or looking at candidates for complex disorders, it is important to determine when a change in DNA sequence is a polymorphism and when it is a mutation. In addition, development of SNP chips for whole genome analysis relies heavily on an understanding of the distribution of polymorphic sites in the genome: location, allele frequencies, population differences in frequencies, and population level linkage disequilibrium information.

Development of the SNP and Mutation databases has relied upon

1. Individual labs with specific interests contributing results from gene studies of patients and normal family members.
2. Completion of draft genome sequences.
3. The decision to develop specific projects to address the question of 'what is normal polymorphism' across populations (HapMap and 1000 genomes).
4. Improvement in sequencing technologies that has increased throughput for the above and allowed sequence data to be deposited

to the public databases and shown against the reference genome for human.

Most **common** Mendelian disorders have been dealt with (because single genes: rare single gene disorders are now being tackled through deep sequencing approaches, see Matt Hurles lectures), but complex genetic disorders, and those with more atypical patterns of inheritance are studied using combinations of genotyping and deep sequencing.
To create and use a SNP chip it is important first to know the bases that are variable in normal populations, and to build up a comprehensive database of all known polymorphisms. Two major projects have worked towards this aim: the 1000 genomes projects and the HapMap projects. HapMap came first, with the aim of genotyping variants in different populations to determine allele frequencies and the extent of linkage disequilibrium at the population level. Information on LD blocks is important when considering the design of SNP chips for population analysis, since haplotypic data means we can infer surrounding SNPs provided the key markers defining the LD block are present on the genotyping chip.

E.g. Assume alleles in LD block in a given population are illustrated by:

Haplotype 1 ACTGCGGT reduced to CCT
Haplotype 2 ACTATAGT reduced to CTT
Haplotype 3 GTCGCGAC reduced to TCC

At position 1: C defines ACT, T defines GTC
At position 2: C defines GCG, T defines ATA
At position 3: T defines GT, C defines AC

The 1000 genomes project adds to the above by re-sequencing genomes from individuals and family trios. The new sequence provides additional SNP data, and information about the rate at which novel mutations arise in the genome through germ-line mutations. Statements from the project home page

http://genome.wellcome.ac.uk/doc_WTX047611.html

- “Any two humans are more than 99 per cent the same at the genetic level: the small fraction of genetic material that varies among people can help to explain individual differences in susceptibility to disease, response to drugs or reaction to environmental factors.”
- “Among the populations whose DNA will be sequenced in the 1000 Genomes Project are: Yoruba in Ibadan, Nigeria; Japanese in Tokyo; Han Chinese in Beijing; Utah residents with ancestry from northern and western Europe; Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Toscani in Italy; Gujarati Indians in Houston; Chinese in metropolitan Denver; people of Mexican ancestry in Los Angeles; and people of African ancestry in the south-western United States. ***These people will be anonymous and will not have any medical information collected on them***.”

Surprisingly, the results from the 1000 genomes project show that each individual carries 250-300 mutations in genes, with 50-100 of these in genes already known to be important in a range of disease phenotypes. This suggests that for most genes, loss of one ‘normal’ copy may not significantly impact on the normal phenotype, i.e., reinforces the idea that haploinsufficiency may not in itself change the phenotype. In fact, most known haploinsufficiency related genes are highly specific in their expression profile, and often have roles in early development of tissues in the embryo.
In addition, we all carry around 60 private changes in our DNA sequence that are unique from the polymorphisms carried by our parents. These arise as germ line mutations.

The ability to carry out genome wide genotyping with SNPs has advanced very rapidly (current commercial chips have 5 million markers). In addition, multiple parallel sequencing approaches make it feasible to re-sequence a whole genome, or at least the coding parts of the genome (the exome) cost effectively (both in terms of time and reagents). These are now attractive

tools for starting to obtain more data on the complex diseases that have been less tractable to prior technologies.

Quantitative v Threshold traits

Quantitative traits are much easier to study, as all individuals can be assigned a value along the continuum. This value can be utilised in the statistical analysis of the genotypic data. In general, we expect the number of contributing alleles to increase in parallel to the genetic contribution to the variance at the quantitative level. Case and control samples are actually ‘extremes’: select individuals from opposite ends of the continuum.

BUT

For a complex trait the affected population is made up from individuals with genetic heterogeneity (i.e. any two individuals are really phenocopies with respect to the trait).

The exact alleles shared at the population level may be more diverse than common alleles shared within a given pedigree. Thus linkage data from multiple pedigrees can be utilised to define chromosomal regions shared between family members, as well as using case-control studies for some genetic traits.

i.e. For complex disorders, where the ‘affected’ population is in reality very genetically heterogeneous, and many controls will share ‘risk’ alleles with the cases, being able to redefine quantitative aspects of the phenotype allows us to pick apart key features using a more precise approach. See more under WTCCC (Wellcome Trust Case Control Consortium) studies.

Using SNP Chips for genome wide studies

The principle behind the SNP chip is very simple. It relies on using probes to the polymorphic sites. Affymetrics chips use a set of matching and mismatched probes for each allele, and the genotype is scored based on an algorithm defined for each matrix of probes. The Illumina chip works on more straightforward single base extension chemistry. Each probe ends a

base before the polymorphic site, and a single nucleotide is added from a ddNTP mix where the bases carry fluorochromes to allow identification of the allele at the scanning step. The use of dideoxynucleotides ensures chain termination immediately after the added base. Each polymorphic site is scored 20-30 times on a chip, so the statistical analysis of the data can be robust.

Although all possible combinations of base changes occur in the genome, C→T and G→A are common, owing to the fact that most CpGs in the genome are methylated, and deamination of methyl-C generates a T. The mismatch can be repaired back to a C on the mutated strand (as the other DNA strand still carries the complementary G) or the base pair can be fully converted to T:A on the next round of DNA replication, so fixing the change in subsequent rounds.

The opposite alleles at a polymorphic site are differently labelled by the dyes. The chip is scanned to produce an image for each fluorescent dye. These images can be artificially superimposed to define the genotype at each allele. Data collected includes the total fluorescence intensity for each channel, and the combined intensity at the specific site on the chip. Data can then be normalised and averaged across the multiple independent assays for that specific SNP.

Each polymorphism is quality controlled based on clustering of the output results: the three possible genotypes will cluster into specific regions on the plot based on total intensity v intensity ratio of the dyes. Note, that by using both parameters, 'odd' results can be included in analyses looking for regions with copy number changes.

SNP chips can detect small regions of amplification or loss of material from regions flanking chromosomal rearrangements and translocations, but cannot detect true reciprocal translocations.

SNP chips are widely applied to

1. Genome wide association studies (GWAS).
2. Copy number variation analysis.

3. Parent of origin effects – looking at parent-child trios to determine the origins of chromosomal regions in phenotypes where an imprinting effect may be suspected.
4. Homozygosity (autozygosity) mapping of single gene disorders with suspected recessive mode of inheritance and founder effects.

Genome wide association analysis is largely used with complex disease. This application is based on the assumption that complex disease is caused by common variants that independently have little impact on phenotype, but in combination may have a sufficiently large impact to push an individual over some threshold that leads to observable disease. Note that genetics and environment may be important, so the genetic contribution to disease may not be 100%. Family studies allow an estimate of how much is contributed by genetics, and this can be expressed as 'heritability'. For example, if a phenotype has 0.8 (or 80%) heritability, this suggests that 80% of the difference (variation) in the phenotype observed in a given population is down to genetics. The remaining 20% is probably 'environmental or other'.

Genome wide association has an advantage over candidate gene approaches. The candidate gene approach is where genes are selected for mutation analysis based on their known roles in pathways or tissues such that it can be hypothesised that mutations are likely to have a phenotypic outcome: e.g., serotonin and dopamine pathways and their related genes have been well studied in depression because it is known that there are perturbations in these pathways that can be modified with drugs that alter, for example, serotonin uptake. Using genome wide association, genes and loci that may not have been considered, because their functions are still unknown, are included in the analysis.

GWAS tends to use case-control statistical analysis. Therefore it is important that the two groups come from the same population (avoid stratification), phenotypic analysis must be rigorous (avoid phenotypes where phenocopy is a possibility- can be difficult where there is heterogeneity of contributing loci in polygenic phenotypes), and if appropriate the age and sex of the subjects must be matched (e.g. the autoimmune disease SLE is more common in women). Alleles tend to be near HWE (Hardy-Weinberg

Equilibrium) in the population as a whole, but can be checked within the sample data.

Association analysis relies on statistical testing to determine changes at the allele frequency between the cases and controls. However, it must be remembered that genotypic frequencies are also important in some disease models.

For two alleles 50% frequency: expect to see 25% AA, 50% Aa, 25% aa

But a shift to 100% genotype Aa would give the same allele frequencies in cases and controls, so the genotypic data showing a shift from HWE may be important in identifying how loci contribute to the phenotype in question, or if they are under selection in a given population sample.

The raw p values generated must also be adjusted to take into consideration multiple testing. The simplest formula is based in dividing the nominal genome wide p values with the number of markers used in the study. More complex algorithms test for association by chance based on random allocation of samples into the ‘case’ or ‘control’ grouping and comparing the results with the ‘true’ case- control values. This helps to reduce detection of false positives in the data set.

Once a region of the genome is identified, then the position of the markers on the genomic map can be used to define genes that could be associated.

NB: as the number of ncRNAs in the genome is increasing with further investigation of intergenic regions, SNPs that cluster in this ‘gene deserts’ cannot be ignored without more investigation of the potential for such gene deserts to either
1. encode a regulatory RNA molecule, or
2. structural motifs important for gene regulation elsewhere.
[The lectures can only touch the surface, but regulatory molecules (or sequences) can act either in trans or in cis to the gene target.]

Raw GWAS outputs can be further mined for information about how the alleles contribute such as are they additive in the genotypic distributions, are multiple identified loci connected to each other in regulatory networks or pathways (which may tell us more about the underlying biology), and what percentage of the genetic contribution is explained by each variant allele at each significant locus.

Animal Studies – GWAS successes

Human influence within a domesticated animal species provides extremes of traits between breeds that are more tractable than human studies, because genetic diversity in any breed will be limited. (In fact, for some breeds you could almost consider individuals to be part of one big pedigree, thus GWAS is both association and linkage!) GWAS can be utilised to identify regions of the genome under selection from the breeding programmes.

In horses, the extremes of size have been studied using GWAS. A few loci have been shown to have large effects, accounting for the majority of the differences observed in modern horses (>80% of the genetic contribution to size can be accounted for by only four loci). By genotyping individuals at these limited loci, it is theoretically possible to predict the approximate size of the animal.

The WTCCC studies

The largest sets of case-controls have been the WTCCC studies: you can choose to follow these up in the literature, rather than spending time discussing them in the lecture. Key points to remember are as follows:

- Large numbers of samples are needed to find regions with small % contributions to the genetic variance.
- Each disease is really a heterogeneous mixture of phenocopies, which will reduce the power of GWAS at any given locus as not all cases have contributions from the same loci.
- Not very successful with the more discontinuous traits where it is hard to define quantitative elements.

- GWAS is hypothesis free, which means all contributions are relevant, not just those in good candidate genes.
- Without a hypothesis the genes identified by GWAS can be a surprise, and many will need following up through functional assessment.
- The SNP could be regulatory, so don't forget the intergenic peaks of association studies are relevant.

Reading list- all are available on line in the cam domain

Original papers

Wellcome Trust Case Control Consortium papers

e.g

Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010, 464:713-20.

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 447:661-78.

Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics. 2008 92:265-72.

And the disease specific papers that have come from the above (see refs within the first paper)

Reviews that address some of the surrounding issues

Maniolo et al, 2009, Finding the missing heritability of complex diseases. Nature 461; 747.

Kong et al, 2009 Parental origin of sequence variants associated with complex diseases. Nature 462:868-74.

Laird, PW, 2010, Principles and challenges of genome wide DNA methylation analysis. Nature Reviews Genet. 11:191-203.

Cirulli and Goldstein, 2010, Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 11:415-25.

Roberts et al, 2010, The genome-wide association study--a new era for common polygenic disorders. J Cardiovasc Transl Res. 2010 3:173-82. Epub 2010 Mar 27.

Wang et al, 2010 Analysing biological pathways in genome wide association studies Nature Reviews Genet 11:843-854.

To follow up on a specific example (or you may want to choose your own)

Herrera et al, 2011, Genetics and epigenetics of obesity Maturitas 69:41-49.

McCarthy, 2010, Genomics, Diabetes and Obesity New England J of Medicine 363:2339-2350.

Teran-Garcia & Bouchard, 2007, Genetics of metabolic syndrome Appl. Physiol. Nutr. Metab. 32:89-114.

Making the most of GWAS requires some novel approaches.

A. Can the phenotype be simplified by looking for more quantitative aspects?
B. Can genes be grouped to spot common themes and pathways?
C. Can we combine different genetic analysis techniques to improve confidence in the output?

Endophenotypes in diabetes combines LOD scores across pedigrees with affected sib pair analysis, so in this example the study is using **linkage** and **not association.** Intervals defined in this way are many Mb, as recombination of parental alleles has not occurred very often: linkage is across large segments of the chromosome. ***The analysis has also used microsats, and not SNPs, but the principle of scanning the whole genome is the same whatever marker is employed***.

Endophenotypes are also useful for association studies at the population level, as we can place the whole population on the continuum, and we can even split our cases by their position on the continuum. Looking at endophenotypes in populations (case-control studies) rather than pedigrees gives finer mapping detail, since it relies on association and LD blocks over shorter physical ranges.

For an example of how this can be used, see the following- included as it shows how more imagination with dissecting the endophenotypes in a complex set of diseases can be productive

Goodbourn PT, Bosten JM, Bargary G, Hogg RE, Lawrance-Owen AJ, Mollon JD. Variants in the 1q21 risk region are associated with a visual

endophenotype of autism and schizophrenia. Genes Brain Behav. 2013 Oct 23.

Using Transancestral haplotypes

In humans, the outbred nature of populations means that it can be hard to identify specific alleles in common across all populations. However, the idea of 'transancestral haplotypes' has been put forward to explain the association of some autoimmune disorders with the MHC, and the principle is similar to that employed in animal genetics in that it looks for the smallest region in common between different populations following a GWAS study, BUT, the alleles involved in the identified haplotype may be subtly different, since the polymorphic variants in each population are not necessarily identical (owing to genetic drift over time). In principle though, the LD blocks in association should be from the same chromosomal intervals (think of this as analogous to defining linkage in pedigrees- the alleles are not identical, but the loci defining the interval are in common). (*See Kaufman KM, et al., Fine mapping of Xq28: both MECP2 and IRAK1 contribute to risk for systemic lupus erythematosus in multiple ancestral groups. Ann Rheum Dis. 2013 Mar;72(3):437-44.)*

The take home message from these studies is that there are many loci, with small effects, therefore we need larger and larger sample sizes to reach statistical significance. The Odds Ratio scores for many implicated loci are very modest (2<1), suggesting small risk contributions to the phenotype.

Is this because the cases are still too heterogeneous? Are we missing something critical in our original hypothesis?

Missing Heritability

With all of the GWAS studies so far, the biggest question is why the alleles identified seem to represent such a small percentage of the genetic variance.

GWAS only looks at alleles, and is thus tied to DNA sequence. Some 'mutations' could be epigenetic (chromatin or DNA modification) ***(carry***

out genome wide histone modification assays, or bisulphite conversion followed by arrays to detect which CpGs are methylated).

GWAS alone does not indicate if there is a parent-of-origin effect (similar to traditional imprinted loci), as it does not tell us about the parental origin of the locus. ***More work is needed using trios to establish parental origins and the link between expression of a specific polymorphic variant and potentially imprinted regions of the genome.***

CNVs arising in the germline could be analogous to LOH in tumour samples: loss of a functional allele exposes a mutation on the homologous chromosome. Alternatively, a few loci are genuinely subject to very tight regulation such that over- or under-expression of the gene product results in a phenotypic outcome. Many of these genes have roles in developmental pathways: recent papers have described CNVs in neurodevelopmental genes, contributing to autism or psychotic phenotypes. Common gene deletion syndromes with developmental phenotypes often leave one allele intact, but show features of haploinsufficiency (see chr 22 deletion syndrome http://omim.org/entry/611867 and DiGeorge syndrome http://omim.org/entry /188400 / VFCS http://omim.org/entry/192430).

Some families in the general population may have large genetic contributions from a small number of highly penetrant loci (*look as though they are close to monogenic with smaller impacts from other loci*): allelic heterogeneity could result in more common mutations having smaller contributions in the general population. ***Suggested approach is to isolate individuals from extreme families where there is strong genetic evidence, and re-sequence the genomes to find the causative mutations***.

Do we understand the gene-gene interactions sufficiently well to predict the contribution of epistasis? Even the ‘single gene disorders’ can show significant variability within the co-morbidities found in the disease. This suggests that the patient’s genetic background has both modifiers and genes that exacerbate features of the phenotype.

Epistasis with coat colour variation is a good example of how gene-gene interactions can modify the outcome. All the known modifiers can be placed into the pathways for development of the melanosomes and the delivery of pigment to the skin/hair. Some of these mutations only exist in the heterozygous state: coat colour mutations in horses can be homozygous lethal (often with phenotypes analogous/homologous to some rare human diseases caused by recessive single gene defects: see the paper referenced on the slide for more information)

Again, this demonstrates how improved understanding **of biological pathways and processes** may further illuminate the impact of genetic variants in the future.

Finally, although SNP chips are used largely to compare allele frequencies, the output can be utilised to identify loss/gain of genetic material by comparing the intensity of the signal on the chip, and to utilise the absolute genotypes when analysing regions for homozygosity with suspected founder effects (recessive traits) or shared expanded haplotypes (dominant traits with suspected founder effect). With trio analysis (parents and child), it is now possible to use combined linkage and association data to determine if there is a parental-origin effect of an associated allele, which allows reinterpretation of the population based data.

For the future, associations with drug treatments are much stronger than other genetic analysis (most alleles increase risk moderately, with odds ratios of 1-2; for impacts of drug metabolism those odds ratios rise above 2). This could allow much better regulation of drug treatment, and prediction of which individuals are more likely to suffer from the side effects.

There are thousands of articles on the topic of personalised medicine/ pharmacogenomics. These are suggestions to give you a flavour, and you may want to look at the Nature Reviews/ Current Opinion series of reviews to find your own examples.

Smith DE, Clémençon B, Hediger MA. Proton-coupled oligopeptide transporter family SLC15: physiological, pharmacological and pathological implications. Mol Aspects Med. 2013 Apr-Jun;34(2-3):323-36.
Maliepaard M, et al. Pharmacogenetics in the evaluation of new drugs: a multiregional regulatory perspective. Nat Rev Drug Discov. 2013 Feb;12(2):103-15.
Joseph PG, Pare G, Ross S, Roberts R, Anand SS. Pharmacogenetics in Cardiovascular Disease: The Challenge of Moving From Promise to Realization: Concepts Discussed at the Canadian Network and Centre for Trials Internationally Network Conference (CANNeCTIN), June 2009. Clin Cardiol. 2013 Sep 17. (*book chapter, but available through @cam domain*).

Other confounding effects

Fortuitous association: this can occur if mutations just happen to arise more often on chromosomes with a specific LD haplotype than on chromosomes with other haplotypes, such that there is a spurious increase in the allele frequency in the cases over the controls. i.e GWAS can throw up false positives, as well as false negatives.

Some case studies

Mapping complex disease loci is not always as simple as it may first appear. Confounding effects to bear in mind include the issues touched on above: e.g. locus heterogeneity (single gene disorders, but different genes in different pedigrees) where phenocopies (same phenotypic outcome) of the disease are found in the population, or penetrance (carrying the disease causing allele increases the risk of developing the disease, but is not sufficient for full expression), where not all carriers of the mutation express the phenotype.

The first case studies are examples of diseases that display these characteristics. Both have tumour related phenotypes, both showed genetic linkage in pedigrees to chromosome 9, and both were studied in Cambridge.

Tuberous sclerosis has been described in the literature since at least the 1960s, and MSSE (Ferguson-Smith disease, named after the grandfather of our current Genetics Professor, Anne Ferguson-Smith) has been described in the literature since the 1930s. Both were investigated using genetic approaches during the era that polymorphic markers became easier to utilise through PCR (the first results were all based on radioactive hybridisation of probes to restriction digested DNA on Southern blots - the restriction fragment length polymorphism approach), but before the release of a reference human genome. MSSE has only recently been solved by re-sequencing patient DNA samples from different populations.

Tuberous sclerosis (techniques: *pedigree based genetic study, linkage analysis using LODs, LOH analysis using MS, candidate gene isolation, re-sequencing of positional candidates for mutations*)

- 1/3 of cases appear to be genetic, as parents and grandparents show mild symptoms, usually the skin features or small fibromas that occur on the nails.
- Autosomal dominant, but reduced penetrance (i.e. <100%).
- Families are only recognised when a severely affected individual is identified, usually at an early age. The patients often have epileptic seizures, and IQ<30.
- Tuberous growths are common in the brain, and in the kidney, but the disease is seen in other tissues.
- Work in the UK showed that there was weak linkage in families to the ABO blood group (gene on 9q). However, with a larger number of pedigrees collected across the world this did not appear to withstand meta-analyses of all of the LOD scores.

Is there more than one gene involved? Is TSC a collection of diseases showing phenocopy?

Some patients with other genetic diseases, including polycystic kidney disease, also had symptoms of TSC. Since the former maps to chromosome 16, this provided further evidence for at least two different loci (chrs 9 and 16).

Pedigrees with confirmed linkage to ABO blood group markers were assembled and separated from families which did not. As the tubers looked a bit like tumours, and this was an era when tumours had been analysed for loss of heterozygosity to find deletion, a similar genome-wide microsatellite scan was performed on patient DNA - matched normal and tuberous tissue. Deletion on chromosome 16 was more frequent than deletion on chromosome 9, but both loci showed LOH in a subset of patients. The regions of LOH were cloned, analysed for gene content and sequenced. This confirmed the identity of two different genes that were involved both in the pedigrees and in sporadic cases of TSC.

Note that this part of chr 16 shows recurrent LOH. This is a genomic region where germ-line loss of chromosomal material results in a high percentage of *de novo* disease.

The proteins encoded by these genes (tuberin and hamartin) function together in the mTORC1 pathway. This pathway is currently under extensive scrutiny for its role in other forms of neurological malfunction, such as epilepsy (common in TSC patients) and autism.

See reviews by authors such as JRW Yates, and papers referenced on slides.

MSSE (multiple, self-healing, squamous epithelioma) (*techniques: pedigree based linkage and haplotype analysis with multiple functional and positional candidates, followed by re-sequencing of interval for detection of mutations)*

- Autosomal dominant, self- healing tumour.
- Huge range in age-of-onset (6 to 60+)
- Tumours tend to occur on exposed areas of the skin, suggesting a role for UV as an environmental trigger.
- Reduced penetrance.

The variation in age-of-onset combined with reduced penetrance made it difficult to identify ‘unaffected’ family members. Many of the collected pedigrees came from coal mining areas of the UK, although other families

were identified in the US, Canada and Australia. A breakthough in the study came when it was noticed that some of the families had similar surnames (but corrupted, so not identical) and genealogists showed that many of the smaller pedigrees were related to a family originating from Ayrshire. There was a strong chance the disease had a founder in the 1700s (coal mining a theme, so perhaps mutation arose in response to environmental factors?) and the extant families should share a common fragment of the genome inherited from this ancestor.

Approach

- Used a pedigree based linkage with microsatellites.
- Developed new markers alongside recombination (genetic) maps of region.
- Looked for the smallest shared haplotype of alleles in all affected individuals – as patients were heterozygous, this required some knowledge about phase of inheritance of the alleles over multiple generations (3 or more) (LOH not discovered in families analysed).
- Knew that chromosome 9 was an important place to search (ABO blood group, again).

Across this extended pedigree, a genetic interval was discovered where the affected individuals shared the same alleles. This haplotype was thought to extend over ~2cM of the chromosome, still a large interval.
It was only recently that the data were revisited. Several other families around the world had been identified, and although they were not related to the original families, they also showed linkage to the same part of chromosome 9. The physical maps and reference human genomic sequences were available allowing a sequencing approach to identify which gene is responsible.
The physical interval is about 4-4.5 Mb and is complicated by the fact that it has multiple, good positional candidate genes: *XPA, FCC, PTCH1.*(Early studies in the lab focussed on resequencing the candidates, with no obvious mutation found in the families). In fact, a gene right at the edge of the interval was mutated, and not the prior candidates. *TGFRB1* is already known for its involvement in Marfan Syndrome, but in MSSE it acts like an

oncogene. The spectrum of mutation is different between the diseases. In Marfan's the changes tend to be amino acid substitutions in the kinase domain. In MSSE, there are terminations in the kinase domain, but also changes to the sequence in the receptor domain. (*TGFRB1* shows **allelic heterogeneity with a pleiotropic outcome**. Contrast with mutations in *CFTR* where there is **allelic heterogeneity with the same disease outcome**.)

Multiple Sclerosis (MS) – an example of GWAS where outputs have been used to identify the pathways contributing to pathogenesis through data-mining and knowledge of biological pathways.
MS is an example of a complex disease with both a genetic and an environmental element. The environmental impact has been confirmed by looking at incidence across the world, and what happens to the incidence when individuals leave an 'ancestral' environment through migration studies. One of the environmental factors is thought to be daylight, or more precisely, vitamin D production (still controversial with some papers claiming association, and others none- also some papers claim the vit D effect is important during pregnancy –i.e. it could be an *in utero* factor!).

However, there is also a strong genetic impact on development of MS, as risk is increased the closer the affected relatives are in the pedigree.
The nature of MS had been debated almost since it was first described: is it driven by neuro-degenerative factors, or by inflammation? Patients definitely showed features consistent with autoimmune disorders, but is this a primary or secondary feature? GWAS outputs can be combined to ask which pathways have largest contribution to disease processes by grouping the associated genes by function. (The principle behind 'systems biology' approaches: a combination of biology and experimental outputs, with computational analysis and data-mining.) In MS the overwhelming answer is that this appears to be a disease promoted by inflammatory responses. In particular, T helper cells are compromised. The GWAS associations have been grouped according to the significant functional processes where genes

are part of the same pathways. Remission can be achieved through targeting with antibody based therapies, as well as interferons.

For more about the pathogenesis and associated MS papers see the references in

http://omim.org/entry/126200

This handout is a short guide, which I hope will help you with your revision. Do refer to the papers and the textbooks for further information, and don't forget to search out your own example of a complex disease that demonstrates the challenges associated with understanding its genetics.

Recommended:

Strachan & Read, ***Human Molecular Genetics***

Strachan, Goodship & Chinery, ***Genetics and Genomics in Medicine***